Queueing Syst (2010) 65: 135–156
DOI 10.1007/s11134-010-9174-1

Interacting queues in heavy traffic

J.A. Morrison · S.C. Borst

Received: 27 May 2009 / Revised: 11 December 2009 / Published online: 21 April 2010

Abstract We consider a system of parallel queues with Poisson arrivals and exponentially distributed service requirements. The various queues are coupled through their service rates, causing a complex dynamic interaction. Specifically, the system consists of one primary queue and several secondary queues whose service rates depend on whether the primary queue is empty or not. Conversely, the service rate of the primary queue depends on which of the secondary queues are empty.

An important special case arises when the service rates satisfy a specific relationship so that the various queues form a work-conserving system. This case is, in fact, equivalent to a so-called Generalized Processor Sharing (GPS) system. GPS-based scheduling algorithms have emerged as popular mechanisms for achieving service differentiation while providing statistical multiplexing gains.

We consider a heavy-traffic scenario, and assume that each of the secondary queues is underloaded when the primary queue is busy. Using a perturbation procedure, we derive the lowest-order asymptotic approximation to the joint stationary distribution of the queue lengths, in terms of a small positive parameter measuring the closeness of the system to instability. Heuristic derivations of these results are presented.

We also pursue two extensions: (i) the more general work-conserving case where the service rate of a secondary queue may depend on its own length, and is a slowly varying function of the length of the primary queue; and (ii) the non-work-conserving case where the service rate of a secondary queue may depend on its own length, but not on the length of the primary queue.

J.A. Morrison is a consultant.
S.C. Borst is also affiliated with Eindhoven University of Technology.

J.A. Morrison · S.C. Borst (✉)
Bell Laboratories, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
e-mail: sem@alcatel-lucent.com

J.A. Morrison
e-mail: johnmorrison@alcatel-lucent.com

Keywords Asymptotic expansion · Bandwidth sharing · Coupled processors · First-order approximation · Heavy traffic · Generalized processor sharing · Perturbation · State-dependent service rates · Time scale decomposition

Mathematics Subject Classification (2000) 60K30 · 90B22

1 Introduction

We consider a system of parallel queues with Poisson arrivals and exponentially distributed service requirements. The various queues are coupled through their service rates, causing a complex dynamic interaction. Specifically, the system consists of one primary queue and several secondary queues whose service rates depend on whether the primary queue is empty or not. Conversely, the service rate of the primary queue depends on which of the secondary queues are empty.

An important special case arises when the service rates satisfy a specific relationship so that the various queues form a work-conserving system. This case is, in fact, equivalent to a so-called Generalized Processor Sharing (GPS) system. GPS-based scheduling algorithms provide effective mechanisms for achieving service differentiation while obtaining statistical multiplexing gains, and have attracted significant interest over the last several years [7, 14, 17, 18]. Large-buffer asymptotics for GPS systems have been obtained for light-tailed [1, 15, 21, 22] and heavy-tailed [2, 3] traffic processes as well as heterogeneous scenarios [2, 4]. Heavy-traffic results may be found in [19, 20].

In the two-queue case, i.e., a single secondary queue, the model under consideration corresponds to the so-called coupled-processors model, as originally investigated by Fayolle and Iasnogorodski [8]. They showed that the solution of the functional equation for the generating function of the joint stationary distribution of the queue lengths may be reduced to a Dirichlet problem in the work-conserving case, and to a Riemann–Hilbert problem in the general case. Related results were obtained in [5, 9, 13].

The work-conserving two-queue case was also examined more recently by Guillemin and Pinchon [10] who obtained a more compact solution of the functional equation based on a Dirichlet problem formulation. The diffusion approximation for the work-conserving case, when both queues are in heavy traffic, was investigated by Morrison [16] who derived explicit integral representations for the solution. Later, Knessl and Morrison [12] extended the heavy-traffic diffusion approximation to the general case, and showed that it is necessary to distinguish nine different regions of the parameter space.

In the present paper, we consider the heavy-traffic behavior with an arbitrary number of secondary queues, assuming that each of these is underloaded when the primary queue is busy. Using a perturbation procedure, we derive the lowest-order asymptotic approximation to the joint stationary distribution of the queue lengths in terms of a small positive parameter measuring the closeness of the system to instability. To this order the various queue lengths are independent, and the secondary queues and the primary queue have geometrically and, after suitable scaling, exponentially distributed lengths, respectively. Heuristic derivations of these results are presented.

We also pursue two extensions: (i) the more general work-conserving case where the service rate of a secondary queue may depend on its own length, and is a slowly varying function of the length of the primary queue; and (ii) the non-work-conserving case where the service rate of a secondary queue may depend on its own length, but not on the length of the primary queue.

The remainder of the paper is organized as follows. In Sect. 2 we formulate the problem and derive the lowest-order asymptotic approximation to the joint stationary distribution of the queue lengths. In Sect. 3 we consider the work-conserving case. An interpretation of the results is discussed in Sect. 4, with some of the supporting technical arguments deferred to Sect. 7. We further examine two extensions: (i) the more general work-conserving case where the service rate of a secondary queue may depend on its own length, and is a slowly varying function of the length of the primary queue; and (ii) the non-work-conserving case where the service rate of a secondary queue may depend on its own length, but not on the length of the primary queue, which are investigated in Sects. 5 and 6, respectively. We make some concluding remarks in Sect. 8.

2 Model formulation and analysis

We consider a system of parallel infinite-buffer queues, with one primary queue, labeled 0, and K secondary queues, indexed by $\mathcal{I} = \{1, \ldots, K\}$. Customers arrive at queue i as a Poisson process of rate λ_i, $i = 0, 1, \ldots, K$, and have exponentially distributed service requirements. Let n_i be the number of customers in queue i, $i = 0, 1, \ldots, K$, $\mathbf{n} = (n_1, \ldots, n_K)$, and $E(\mathbf{n}) := \{j \in \mathcal{I} : n_j = 0\}$ the set of empty secondary queues. The service rate for queue i is μ_i when queue 0 is busy, and $\nu_i(E(\mathbf{n}))$ otherwise, $i = 1, \ldots, K$. We assume that $\lambda_i < \mu_i$ for all $i \in \mathcal{I}$, i.e., each of the secondary queues is underloaded when the primary queue is busy. The service rate for queue 0 is $\mu_0 + \sum_{i \in E(\mathbf{n})} \kappa_i = \mu_0 + \sum_{i=1}^{K} \mathrm{I}_{\{n_i = 0\}} \kappa_i$, where μ_0 and κ_i, $i = 1, \ldots, K$, are constants independent of $n_0 \neq 0$ and $\mathbf{n}$. We assume that $\mu_0 + \sum_{i=1}^{K} \mathrm{I}_{\{\kappa_i < 0\}} \kappa_i \geq 0$. We will assume that the primary queue is heavily loaded, i.e.,

$$\lambda_0(1 + \delta) = \mu_0 + \sum_{i=1}^{K} \kappa_i \left(1 - \frac{\lambda_i}{\mu_i}\right), \tag{1}$$

with $0 < \delta \ll 1$.

Let $N(t) := (N_0(t), N_1(t), \ldots, N_K(t))$ be the Markov process representing the number of customers in the various queues at time t. The system is said to be stable if the process $N(t)$ has a stationary distribution. Establishing necessary and sufficient conditions for stability is difficult in general and outside the scope of the present paper. For δ sufficiently small, Condition (1) guarantees stability in the case $\kappa_i \geq 0$ for all i with $\lambda_i \geq \min_{E \subseteq \{1, \ldots, N\}} \nu_i(E)$. This may be shown by considering the corresponding fluid limit [6]. In particular, Condition (1) is sufficient to ensure stability in the case $\kappa_i \geq 0$ for all $i = 1, \ldots, K$.

Assuming the above condition holds, let $p(n_0, \mathbf{n}) := \lim_{t \to \infty} \mathbb{P}\{N(t) = (n_0, \mathbf{n})\}$ be the stationary probability that there are n_i customers in queue i, $i = 0, 1, \ldots, K$.

Then the balance equations satisfied by $p(n_0, \mathbf{n})$ are

$$\begin{aligned}
&\left\{\lambda_0 + \mu_0 + \sum_{i=1}^{K}[\lambda_i + \mathrm{I}_{\{n_i=0\}}\kappa_i + \mathrm{I}_{\{n_i>0\}}\mu_i]\right\} p(n_0, \mathbf{n}) \\
&\quad = \lambda_0 p(n_0 - 1, \mathbf{n}) + \left[\mu_0 + \sum_{i=1}^{K} \mathrm{I}_{\{n_i=0\}}\kappa_i\right] p(n_0 + 1, \mathbf{n}) \\
&\qquad + \sum_{i=1}^{K}\big[\lambda_i \mathrm{I}_{\{n_i>0\}} p(n_0, \mathbf{n} - \mathbf{e}_i) + \mu_i p(n_0, \mathbf{n} + \mathbf{e}_i)\big], \quad n_0 \geq 1, \qquad (2) \\
&\left\{\lambda_0 + \sum_{i=1}^{K}\big[\lambda_i + \mathrm{I}_{\{n_i>0\}}\nu_i\big(E(\mathbf{n})\big)\big]\right\} p(0, \mathbf{n}) \\
&\quad = \left[\mu_0 + \sum_{i=1}^{K} \mathrm{I}_{\{n_i=0\}}\kappa_i\right] p(1, \mathbf{n}) \\
&\qquad + \sum_{i=1}^{K}\big[\lambda_i \mathrm{I}_{\{n_i>0\}} p(0, \mathbf{n} - \mathbf{e}_i) + \nu_i\big(E(\mathbf{n} + \mathbf{e}_i)\big) p(0, \mathbf{n} + \mathbf{e}_i)\big], \qquad (3)
\end{aligned}$$

with $\mathbf{e}_i$ denoting the ith unit vector.

We let $\xi = \delta n_0$ and

$$p(\xi/\delta, \mathbf{n}) = \delta\phi(\xi, \mathbf{n}; \delta), \quad 0 < \xi = \mathrm{O}(1). \qquad (4)$$

Then, from (1), since $p(n_0 \pm 1, \mathbf{n}) = \delta\phi(\xi \pm \delta, \mathbf{n}; \delta)$, a Taylor series expansion yields

$$\begin{aligned}
&\left\{\lambda_0 + \mu_0 + \sum_{i=1}^{K}[\lambda_i + \mathrm{I}_{\{n_i=0\}}\kappa_i + \mathrm{I}_{\{n_i>0\}}\mu_i]\right\} \phi(\xi, \mathbf{n}; \delta) \\
&\quad = \lambda_0\left[\phi(\xi, \mathbf{n}; \delta) - \delta\frac{\partial\phi}{\partial\xi}(\xi, \mathbf{n}; \delta) + \frac{\delta^2}{2}\frac{\partial^2\phi}{\partial\xi^2}(\xi, \mathbf{n}; \delta) + \mathrm{O}(\delta^3)\right] \\
&\qquad + \left[\mu_0 + \sum_{i=1}^{K} \mathrm{I}_{\{n_i=0\}}\kappa_i\right]\left[\phi(\xi, \mathbf{n}; \delta) + \delta\frac{\partial\phi}{\partial\xi}(\xi, \mathbf{n}; \delta)\right. \\
&\qquad \left. + \frac{\delta^2}{2}\frac{\partial^2\phi}{\partial\xi^2}(\xi, \mathbf{n}; \delta) + \mathrm{O}\big(\delta^3\big)\right] \\
&\qquad + \sum_{i=1}^{K}\big[\lambda_i \mathrm{I}_{\{n_i>0\}}\phi(\xi, \mathbf{n} - \mathbf{e}_i; \delta) + \mu_i\phi(\xi, \mathbf{n} + \mathbf{e}_i; \delta)\big], \qquad (5)
\end{aligned}$$

which, after some rearrangement and use of (1), implies that

$$\sum_{i=1}^{K}\{\big[\lambda_i\phi(\xi,\mathbf{n};\delta)-\mu_i\phi(\xi,\mathbf{n}+\mathbf{e}_i;\delta)\big]-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\phi(\xi,\mathbf{n}-\mathbf{e}_i;\delta)-\mu_i\phi(\xi,\mathbf{n};\delta)\big]\}$$

$$=\delta\left\{\sum_{i=1}^{K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\kappa_i+\delta\lambda_0\right\}\frac{\partial\phi}{\partial\xi}(\xi,\mathbf{n};\delta)$$

$$+\frac{\delta^2}{2}\left\{2\lambda_0+\sum_{i=1}^{K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\kappa_i\right\}\frac{\partial^2\phi}{\partial\xi^2}(\xi,\mathbf{n};\delta)+\mathrm{O}\big(\delta^3\big). \tag{6}$$

If we sum over $\mathbf{n}=(n_1,\ldots,n_K)$, we obtain

$$\sum_{i=1}^{K}\kappa_i\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\frac{\partial\phi}{\partial\xi}(\xi,\mathbf{n};\delta)$$

$$+\delta\lambda_0\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\partial\phi}{\partial\xi}(\xi,\mathbf{n};\delta)+\frac{\partial^2\phi}{\partial\xi^2}(\xi,\mathbf{n};\delta)\right]$$

$$+\frac{\delta}{2}\sum_{i=1}^{K}\kappa_i\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\frac{\partial^2\phi}{\partial\xi^2}(\xi,\mathbf{n};\delta)+\mathrm{O}\big(\delta^2\big)=0. \tag{7}$$

We expand in powers of δ, and let

$$\phi(\xi,\mathbf{n};\delta)=\phi^{(0)}(\xi,\mathbf{n})+\delta\phi^{(1)}(\xi,\mathbf{n})+\mathrm{O}\big(\delta^2\big), \tag{8}$$

so to lowest order, from (6), we obtain

$$\sum_{i=1}^{K}\{\big[\lambda_i\phi^{(0)}(\xi,\mathbf{n})-\mu_i\phi^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i)\big]$$

$$-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\phi^{(0)}(\xi,\mathbf{n}-\mathbf{e}_i)-\mu_i\phi^{(0)}(\xi,\mathbf{n})\big]\}=0. \tag{9}$$

We let $\phi^{(0)}(\xi,\mathbf{n})=\pi^{(0)}(\xi,\mathbf{n})P_0(\xi)$, where $\sum_{\mathbf{n}\in\mathbf{N}^K}\pi^{(0)}(\xi,\mathbf{n})=1$. Then the conditional probability $\pi^{(0)}(\xi,\mathbf{n})$ satisfies the same equation as $\phi^{(0)}(\xi,\mathbf{n})$. We note that $\delta(n_0\pm1)=\xi\pm\delta$ so that ξ changes by order δ when there is an arrival to, or departure from, queue 0. There is a solution

$$\pi^{(0)}(\xi,\mathbf{n})=\prod_{i=1}^{K}\left(1-\frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}. \tag{10}$$

Since $\lambda_i\pi^{(0)}(\xi,\mathbf{n})=\mu_i\pi^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i)$, $i=1,\ldots,K$, this is the stationary distribution according to Theorem 1.3 in [11], and

$$\phi^{(0)}(\xi,\mathbf{n})=\prod_{i=1}^{K}\left(1-\frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}P_0(\xi), \tag{11}$$

where $P_0(\xi)$ is yet to be determined.

Next, from (6) and (8), we obtain

$$\begin{aligned}&\sum_{i=1}^{K}\Big\{\big[\lambda_i\phi^{(1)}(\xi,\mathbf{n})-\mu_i\phi^{(1)}(\xi,\mathbf{n}+\mathbf{e}_i)\big]\\&\qquad-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\phi^{(1)}(\mathbf{n}-\mathbf{e}_i)-\mu_i\phi^{(1)}(\xi,\mathbf{n})\big]\Big\}\\&\quad=\sum_{i=1}^{K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\kappa_i\frac{\partial\phi^{(0)}}{\partial\xi}(\xi,\mathbf{n}).\end{aligned}\tag{12}$$

We sum over n_j, $j\neq l$, and let

$$m_l^{(1)}(\xi,n_l)=\sum_{j\neq l}\sum_{n_j=0}^{\infty}\phi^{(1)}(\xi,\mathbf{n}).\tag{13}$$

We note, from (11), that

$$\sum_{n_j=0}^{\infty}\mathrm{I}_{\{n_j>0\}}\phi^{(0)}(\xi,\mathbf{n})=\frac{\lambda_j}{\mu_j}\sum_{n_j=0}^{\infty}\phi^{(0)}(\xi,\mathbf{n}).\tag{14}$$

Hence, from (12)–(14), we obtain

$$\begin{aligned}&\lambda_l m_l^{(1)}(\xi,n_l)-\mu_l m_l^{(1)}(\xi,n_l+1)\\&\qquad-\mathrm{I}_{\{n_l>0\}}\big[\lambda_l m_l^{(1)}(\xi,n_l-1)-\mu_l m_l^{(1)}(\xi,n_l)\big]\\&\quad=\left[\frac{\lambda_l}{\mu_l}-\mathrm{I}_{\{n_l>0\}}\right]\kappa_l\left(1-\frac{\lambda_l}{\mu_l}\right)\left(\frac{\lambda_l}{\mu_l}\right)^{n_l}\frac{\mathrm{d}P_0}{\mathrm{d}\xi}.\end{aligned}\tag{15}$$

If we sum over n_l from 0 to r, we obtain

$$\lambda_l m_l^{(1)}(\xi,r)-\mu_l m_l^{(1)}(\xi,r+1)=\kappa_l\left(1-\frac{\lambda_l}{\mu_l}\right)\left(\frac{\lambda_l}{\mu_l}\right)^{r+1}\frac{\mathrm{d}P_0}{\mathrm{d}\xi}.\tag{16}$$

But, from (14),

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\phi^{(0)}(\xi,\mathbf{n})=0.\tag{17}$$

Hence the O(1) terms in (7) cancel, and we obtain

$$\begin{aligned}&\sum_{i=1}^{K}\kappa_i\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\frac{\partial\phi^{(1)}}{\partial\xi}(\xi,\mathbf{n})\\&\qquad+\lambda_0\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\partial\phi^{(0)}}{\partial\xi}(\xi,\mathbf{n})+\frac{\partial^2\phi^{(0)}}{\partial\xi^2}(\xi,\mathbf{n})\right]=0.\end{aligned}\tag{18}$$

But, from (13) and (16),

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\phi^{(1)}(\xi,\mathbf{n})=\sum_{n_i\geq 0}\left[\frac{\lambda_i}{\mu_i}m_i^{(1)}(\xi,n_i)-m_i^{(1)}(\xi,n_i+1)\right]$$
$$=\frac{\kappa_i\lambda_i}{\mu_i^2}\frac{\mathrm{d}P_0}{\mathrm{d}\xi}. \tag{19}$$

Hence, from (11), (18) and (19), we obtain

$$\frac{\mathrm{d}P_0}{\mathrm{d}\xi}+\left(1+\frac{1}{\lambda_0}\sum_{i=1}^{K}\frac{\kappa_i^2\lambda_i}{\mu_i^2}\right)\frac{\mathrm{d}^2P_0}{\mathrm{d}\xi^2}=0. \tag{20}$$

From the normalization condition, $\sum_{n_0=0}^{\infty}\sum_{\mathbf{n}\in\mathbf{N}^K}p(n_0,\mathbf{n})=1$, and the Euler–Maclaurin summation formula, $\int_0^{\infty}P_0(\xi)\,\mathrm{d}\xi=1$. Hence,

$$P_0(\xi)=\tau\mathrm{e}^{-\tau\xi}, \tag{21}$$

where

$$\tau^{-1}=1+\frac{1}{\lambda_0}\sum_{i=1}^{K}\frac{\lambda_i\kappa_i^2}{\mu_i^2}. \tag{22}$$

Let

$$\hat{\phi}^{(1)}(\xi,\mathbf{n})=-\prod_{j=1}^{K}\left(1-\frac{\lambda_j}{\mu_j}\right)\left(\frac{\lambda_j}{\mu_j}\right)^{n_j}\sum_{l=1}^{K}\frac{\kappa_l}{\mu_l}n_l\frac{\mathrm{d}P_0}{\mathrm{d}\xi}. \tag{23}$$

Then,

$$\lambda_i\hat{\phi}^{(1)}(\xi,\mathbf{n})-\mu_i\hat{\phi}^{(1)}(\xi,\mathbf{n}+\mathbf{e}_i)=\prod_{j=1}^{K}\left(1-\frac{\lambda_j}{\mu_j}\right)\left(\frac{\lambda_j}{\mu_j}\right)^{n_j}\frac{\lambda_i\kappa_i}{\mu_l}\frac{\mathrm{d}P_0}{\mathrm{d}\xi}, \tag{24}$$

implying

$$\lambda_i\hat{\phi}^{(1)}(\xi,\mathbf{n})-\mu_i\hat{\phi}^{(1)}(\xi,\mathbf{n}+\mathbf{e}_i)-\mathrm{I}_{\{n_i>0\}}\left[\lambda_i\hat{\phi}^{(1)}(\xi,\mathbf{n}-\mathbf{e}_i)-\mu_i\hat{\phi}^{(1)}(\xi,\mathbf{n})\right]$$
$$=\prod_{j=1}^{K}\left(1-\frac{\lambda_j}{\mu_j}\right)\left(\frac{\lambda_j}{\mu_j}\right)^{n_j}\left[\frac{\lambda_i}{\mu_i}-\mathrm{I}_{\{n_i>0\}}\right]\kappa_i\frac{\mathrm{d}P_0}{\mathrm{d}\xi}. \tag{25}$$

Hence, $\hat{\phi}^{(1)}(\xi,n)$ is a particular solution of (12) for $\phi^{(1)}(\xi,\mathbf{n})$. We note that

$$\frac{\hat{\phi}^{(1)}(\xi,\mathbf{n})}{\phi^{(0)}(\xi,\mathbf{n})}=-\sum_{l=1}^{K}\frac{\kappa_l}{\mu_l}n_l\frac{\mathrm{d}P_0}{\mathrm{d}\xi}/P_0(\xi)=\sum_{l=1}^{K}\frac{\kappa_l}{\mu_l}n_l\Big/\left(1+\frac{1}{\lambda_0}\sum_{i=1}^{K}\frac{\lambda_i\kappa_i^2}{\mu_i^2}\right), \tag{26}$$

and the number of terms in both the numerator and the denominator increase linearly as K increases. We also note that $(\frac{\lambda_j}{\mu_j})^{n_j}$ is exponentially small for $n_j \gg 1$, $j = 1, \ldots, K$.

The complete determination of $\phi^{(1)}(\xi, \mathbf{n})$ would involve the analysis to one more order in δ. Also, in our analysis of $\phi^{(0)}(\xi, \mathbf{n})$, the boundary condition (3) for $n_0 = 0$ is not used. Equation (3) would be required in the determination of $p(n_0, \mathbf{n})$ for $n_0 = \mathrm{O}(1)$, which affects the normalization condition for $\phi^{(1)}(\xi, \mathbf{n})$.

3 Generalized processor sharing

We now consider the work-conserving case, in which a single processor works at a unit rate. If queue 0 is busy, then the processor devotes fractions γ_i of its effort to busy queues i, $i = 1, \ldots, K$, and the remaining fraction $1 - \sum_{i=1}^{K} \mathrm{I}_{\{n_i > 0\}} \gamma_i$ to queue 0. Thus if the required exponential service time for queue i has mean $1/\omega_i$, $i = 0, 1, \ldots, K$, then $\mu_i = \gamma_i \omega_i$, $\kappa_i = \gamma_i \omega_0$, $i = 1, \ldots, K$, $\mu_0 = (1 - \sum_{i=1}^{K} \gamma_i)\omega_0$. (If queue 0 is empty, then it is assumed that $\sum_{i=1}^{K} \mathrm{I}_{\{n_i > 0\}} \nu_i(E(\mathbf{n}))/\omega_i = 1$, $E(\mathbf{n}) \neq \mathcal{I}$.) Hence, independently of γ_i, $\kappa_i/\mu_i = \omega_0/\omega_i$, and

$$\tau^{-1} = 1 + \frac{\omega_0^2}{\lambda_0} \sum_{i=1}^{K} \frac{\lambda_i}{\omega_i^2}, \qquad \frac{\lambda_0}{\omega_0}(1+\delta) + \sum_{i=1}^{K} \frac{\lambda_i}{\omega_i} = 1. \tag{27}$$

Since τ is independent of γ_i, it would seem to be advantageous for the processor to devote all of its effort to the K secondary queues when they are not empty, so that $\sum_{i=1}^{K} \gamma_i = 1$ and $\mu_0 = 0$. In the particular case $\gamma_i = \lambda_i/\rho\omega_i$, $i = 1, \ldots, K$, it follows that $\lambda_i/\mu_i = \rho = \sum_{i=1}^{K} \lambda_i/\omega_i < 1$. Hence,

$$\phi^{(0)}(\xi, \mathbf{n}) = (1-\rho)^K \rho^{\sum_{i=1}^{K} n_i} P_0(\xi), \tag{28}$$

and

$$\hat{\phi}^{(1)}(\xi, \mathbf{n}) = -(1-\rho)^K \rho^{\sum_{i=1}^{K} n_i} \omega_0 \sum_{l=1}^{K} \frac{n_l}{\omega_l} \frac{\mathrm{d}P_0}{\mathrm{d}\xi}. \tag{29}$$

We note that $\hat{\phi}^{(1)}(\xi, \mathbf{n})$ is not a function of $\sum_{i=1}^{K} n_i$, unless $\omega_l = \omega$ for all $l = 1, \ldots, K$.

4 Interpretation

We now provide an interpretation of the results obtained in the two previous sections. Formulas (11) and (21) show that in the heavy-traffic regime the following three asymptotic properties hold: (i) The number of customers of each secondary queue i is geometrically distributed with parameter λ_i/μ_i; (ii) The (scaled) number of customers of the primary queue is exponentially distributed with parameter τ (independent of the values of the fractions γ_i, $i = 1, \ldots, K$, in the

work-conserving case); (iii) the number of customers of the various secondary queues and the (scaled) number of customers of the primary queue are independent.

These three observations may be heuristically explained as follows. First of all, the fraction of time that the heavily loaded primary queue is empty is negligible. Thus, each underloaded secondary queue will practically never receive service at rate $\nu_i(E(\mathbf{n}))$ and will be served at rate μ_i virtually all the time. It follows that the values of the $\nu_i(E(\mathbf{n}))$'s are irrelevant, and that queue i effectively behaves as in an isolated system with constant service rate μ_i, and in particular the number of class-i customers will be geometrically distributed with parameter λ_i/μ_i, yielding Property (i).

Turning attention to Property (ii), it is useful to view the system in a slightly different but equivalent way. Class-0 customers have exponentially distributed service requirements with parameter $\zeta := \mu_0 + \sum_{i=1}^{K} \kappa_i$, while class-$i$ customers have exponentially distributed service requirements with parameters $\zeta\mu_i/\kappa_i$. As long as queue 0 is non-empty, the total service rate is unity, with a portion $(\mu_0 + \sum_{i\in E} \kappa_i)/\zeta$ allocated to the primary queue and portions κ_i/ζ allocated to each of the non-empty secondary queues $i \notin E$. As noted above, the fraction of time that queue 0 is empty is asymptotically negligible. Hence, it will make no difference if we assume that the total service rate is also unity and allocated among the non-empty secondary queues in an arbitrary fashion when queue 0 is empty, regardless of the values of the $\nu_i(E(\mathbf{n}))$'s. It follows that the total workload in the system behaves as that in an M/H$_{K+1}$/1 queue with arrival rate $\lambda = \lambda_0 + \sum_{i=1}^{K} \lambda_i$ and service requirements that are exponentially distributed with parameter ζ with probability λ_0/λ, and parameter $\zeta\mu_i/\kappa_i$ with probability λ_i/λ, $i = 1, \ldots, K$. Standard heavy-traffic results for the M/G/1 queue then imply that the total workload, scaled by $\delta\lambda_0/\zeta$, converges to an exponentially distributed random variable with mean $\frac{\lambda_0}{\zeta^2} + \sum_{i=1}^{K} \frac{\lambda_i\kappa_i^2}{\zeta^2\mu_i^2}$, as $\delta \downarrow 0$. Since nearly the entire workload is concentrated in class-0 customers, each with mean service requirement $1/\zeta$, it follows that their number, scaled by δ, is also exponentially distributed with mean

$$\frac{1}{\lambda_0}\left[\lambda_0 + \sum_{i=1}^{K} \frac{\lambda_i\kappa_i^2}{\mu_i^2}\right] = \tau^{-1},$$

as $\delta \downarrow 0$, yielding Property (ii).

An alternate reasoning starts from the observation that class 0 behaves as an M/M/1 queue with arrival rate λ_0, mean service requirement ζ^{-1}, and variable service speed. Specifically, the service speed is $(\mu_0 + \sum_{i\in E} \kappa_i)/\zeta$ when the subset of empty secondary queues is E. Thus the amount of class-0 work at time t may be expressed as $V_0(t) = \sup_{s\le t}\{A_0(s,t) - B_0(s,t)\}$, with $A_0(s,t)$ denoting the amount of class-0 traffic arriving during the time interval $[s,t]$, $B_0(s,t) := (t-s) - \sum_{i=1}^{K} \kappa_i T_i(s,t)/\zeta$ denoting the amount of service provided to class 0 during the time interval $[s,t]$ and $T_i(s,t)$ representing the amount of time that queue i is non-empty during the time interval $[s,t]$. Note that

$$\mathbb{E}\big\{A_0(0,1) - B_0(0,1)\big\} = \lambda_0/\zeta - 1 + \sum_{i=1}^{K} \kappa_i \frac{\lambda_i}{\mu_i} \Big/ \zeta$$

$$= \left(\lambda_0 - \mu_0 - \sum_{i=1}^{K} \kappa_i \left(1 - \frac{\lambda_i}{\mu_i}\right)\right) \Big/ \zeta = -\delta\lambda_0/\zeta.$$

Thus, as $\delta \downarrow 0$, the amount of class-0 work, scaled with $\delta\lambda_0/\zeta$ is exponentially distributed with mean $\lim_{t\to\infty} \frac{1}{t} \operatorname{Var}\{A_0(0,t) - B_0(0,t)\} = \operatorname{Var}\{A_0(0,1)\} + \sum_{i=1}^{K} \kappa_i^2 \lim_{t\to\infty} \frac{1}{t\zeta^2} \operatorname{Var}\{T_i(0,t)\}$. It is easily verified that $\operatorname{Var}\{A_0(0,1)\} = \lambda_0/\zeta^2$. In addition, it may be checked that $\lim_{t\to\infty} \frac{1}{t} \operatorname{Var}\{T_i(0,t)\} = \lambda_i \mathbb{E}\{P_i^2\}/(1+\lambda_i \mathbb{E}\{P_i\})^3$, with P_i representing the busy period of an isolated M/M/1 queue with arrival rate λ_i and service rate μ_i. Since $\mathbb{E}\{P_i\} = 1/(\mu_i - \lambda_i)$ and $\mathbb{E}\{P_i^2\} = 2\mu_i/(\mu_i - \lambda_i)^3$, it follows that $\lim_{t\to\infty} \frac{1}{t} \operatorname{Var}\{T_i(0,t)\} = 2\lambda_i/\mu_i^2$. Thus, the amount of class-0 work scaled with $\delta\lambda_0/\zeta$ is exponentially distributed with mean $\lambda_0/\zeta^2 + \sum_{i=1}^{K} \frac{\lambda_i \kappa_i^2}{\mu_i^2 \zeta^2}$. Because class-0 customers have mean service requirement $1/\zeta$, this implies that their number, scaled with δ, is also exponentially distributed with mean

$$\frac{1}{\lambda_0}\left[\lambda_0 + \sum_{i=1}^{K} \frac{\lambda_i \kappa_i^2}{\mu_i^2}\right] = \tau^{-1}.$$

Finally, Property (iii) basically results from the separation of time scales that is induced by the heavy-traffic regime, where the dynamics of the secondary class evolve on an O(1) time scale, while the dynamics of the primary class occur on an O($1/\delta$) time scale. Since the amount of 'memory' of the secondary classes asymptotically vanishes compared to that of the primary class, the number of secondary customers is, in the limit, independent of the (scaled) number of primary customers.

In order to further support the above heuristic arguments, we provide in Sect. 7 a rigorous proof of the counterparts of Properties (i)–(iii) in terms of workloads rather than numbers of customers.

5 General work-conserving case

We now examine the heavy-traffic behavior in the more general work-conserving case, so that $\frac{\lambda_0}{\omega_0}(1+\delta) + \sum_{i=1}^{K} \frac{\lambda_i}{\omega_i} = 1$, as before, but we now let

$$\mu_i(n_0, n_i) = g_i(\delta n_0, n_i)\omega_i, \qquad \kappa_i = g_i(\delta n_0, n_i)\omega_0, \quad i = 1, \ldots, K, \tag{30}$$

and

$$\mu_0 = \left[1 - \sum_{i=1}^{K} g_i(\delta n_0, n_i)\right]\omega_0 \geq 0.$$

We assume that $g_i(\xi, n_i)\omega_i/\lambda_i \geq 1/\rho_i > 1$, for $\xi > 0$ and $n_i \geq 1$, $i = 1, \ldots, K$, and that $g_i(\xi, n_i)$ is differentiable for $\xi > 0$. The balance equations satisfied by $p(n_0, \mathbf{n})$

are

$$\begin{aligned}
&\left\{\lambda_0+\left[1-\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}g_i(\delta n_0,n_i)\right]\omega_0+\sum_{i=1}^{K}\left[\lambda_i+\mathrm{I}_{\{n_i>0\}}g_i(\delta n_0,n_i)\omega_i\right]\right\}p(n_0,\mathbf{n})\\
&\quad=\lambda_0 p(n_0-1,\mathbf{n})+\omega_0\left[1-\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}g_i(\delta n_0+\delta,n_i)\right]p(n_0+1,\mathbf{n})\\
&\qquad+\sum_{i=1}^{K}\left[\lambda_i\mathrm{I}_{\{n_i>0\}}p(n_0,\mathbf{n}-\mathbf{e}_i)+g_i(\delta n_0,n_i+1)\omega_i p(n_0,\mathbf{n}+\mathbf{e}_i)\right],\\
&\quad n_0\geq 1. \qquad (31)
\end{aligned}$$

We let $\xi=\delta n_0$ and

$$p(\xi/\delta,\mathbf{n})=\delta\chi(\xi,\mathbf{n};\delta),\quad 0<\xi=\mathrm{O}(1). \qquad (32)$$

Then, expanding $\chi(\xi-\delta,\mathbf{n};\delta)$ and

$$\left[1-\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}g_i(\xi+\delta,n_i)\right]\chi(\xi+\delta,\mathbf{n};\delta)$$

in a Taylor series, we obtain

$$\begin{aligned}
&\sum_{i=1}^{K}\left\{\left[\lambda_i+\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\omega_i\right]\chi(\xi,\mathbf{n};\delta)\right.\\
&\qquad\left.-\mathrm{I}_{\{n_i>0\}}\lambda_i\chi(\xi,\mathbf{n}-\mathbf{e}_i;\delta)-g_i(\xi,n_i+1)\omega_i\chi(\xi,\mathbf{n}+\mathbf{e}_i;\delta)\right\}\\
&\quad=\delta\omega_0\frac{\partial}{\partial\xi}\left\{\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi(\xi,\mathbf{n};\delta)\right\}\\
&\qquad+\delta^2\lambda_0\left[\frac{\partial\chi}{\partial\xi}(\xi,\mathbf{n};\delta)+\frac{\partial^2\chi}{\partial\xi^2}(\xi,\mathbf{n};\delta)\right]\\
&\qquad+\frac{\delta^2}{2}\omega_0\frac{\partial^2}{\partial\xi^2}\left\{\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,\mathbf{n})\right]\chi(\xi,\mathbf{n};\delta)\right\}+\mathrm{O}\big(\delta^3\big). \qquad (33)
\end{aligned}$$

Now,

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}\lambda_i\chi(\xi,\mathbf{n}-\mathbf{e}_i;\delta)=\sum_{i=1}^{K}\lambda_i\sum_{\mathbf{n}\in\mathbf{N}^K}\chi(\xi,\mathbf{n};\delta), \qquad (34)$$

and

$$\begin{aligned}
&\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\omega_i\chi(\xi,\mathbf{n};\delta)\\
&\quad=\sum_{i=1}^{K}\sum_{\mathbf{n}\in\mathbf{N}^K}g_i(\xi,n_i+1)\omega_i\chi(\xi,\mathbf{n}+\mathbf{e}_i;\delta). \qquad (35)
\end{aligned}$$

Hence, if we sum over $\mathbf{n}$ in (33), we obtain

$$\begin{aligned}
&\omega_0 \frac{\partial}{\partial \xi}\left\{\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi(\xi,\mathbf{n};\delta)\right\}\\
&\quad+\delta\lambda_0\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\partial\chi}{\partial\xi}(\xi,\mathbf{n};\delta)+\frac{\partial^2\chi}{\partial\xi^2}(\xi,\mathbf{n};\delta)\right]\\
&\quad+\frac{\delta}{2}\omega_0\frac{\partial^2}{\partial\xi^2}\left\{\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi(\xi,\mathbf{n};\delta)\right\}\\
&\quad+\mathrm{O}\big(\delta^2\big)=0. \qquad (36)
\end{aligned}$$

We expand in powers of δ, and let

$$\chi(\xi,\mathbf{n};\delta)=\chi^{(0)}(\xi,\mathbf{n})+\delta\chi^{(1)}(\xi,\mathbf{n})+\mathrm{O}\big(\delta^2\big). \qquad (37)$$

Then, from (33), we obtain

$$\begin{aligned}
&\sum_{i=1}^{K}\big\{\big[\lambda_i\chi^{(0)}(\xi,\mathbf{n})-g_i(\xi,n_i+1)\omega_i\chi^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i)\big]\\
&\quad-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\chi^{(0)}(\xi,\mathbf{n}-\mathbf{e}_i)-g_i(\xi,n_i)\omega_i\chi^{(0)}(\xi,\mathbf{n})\big]\big\}=0. \qquad (38)
\end{aligned}$$

We let $\chi^{(0)}(\xi,\mathbf{n})=M^{(0)}(\xi,\mathbf{n})X_0(\xi)$, where $\sum_{\mathbf{n}\in\mathbf{N}^K}M^{(0)}(\xi,\mathbf{n})=1$. Then the conditional probability $M^{(0)}(\xi,\mathbf{n})$ satisfies the same (38) as $\chi^{(0)}(\xi,\mathbf{n})$, and there is a solution

$$M^{(0)}(\xi,\mathbf{n})=\prod_{j=1}^{K}C_j(\xi)\left(\frac{\lambda_j}{\omega_j}\right)^{n_j}\prod_{k=1}^{n_j}\frac{1}{g_j(\xi,k)}, \qquad (39)$$

where

$$C_j(\xi)=\left[\sum_{n_j=0}^{\infty}\left(\frac{\lambda_j}{\omega_j}\right)^{n_j}\prod_{k=1}^{n_j}\frac{1}{g_j(\xi,k)}\right]^{-1}. \qquad (40)$$

Since $\lambda_i M^{(0)}(\xi,\mathbf{n})=g_i(\xi,n_i+1)\omega_i M^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i)$, $i=1,\ldots,K$, this is the stationary distribution, according to Theorem 1.3 in [11], and

$$\chi^{(0)}(\xi,\mathbf{n})=\prod_{j=1}^{K}C_j(\xi)\left(\frac{\lambda_j}{\omega_j}\right)^{n_j}\prod_{k=1}^{n_j}\frac{1}{g_j(\xi,k)}X_0(\xi), \qquad (41)$$

and $\sum_{\mathbf{n}\in\mathbf{N}^K}\chi^{(0}(\xi,\mathbf{n})=X_0(\xi)$. We further have the identity

$$\sum_{n_i=0}^{\infty}\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\left(\frac{\lambda_i}{\omega_i}\right)^{n_i}\prod_{k=1}^{n_i}\frac{1}{g_i(\xi,k)}=\frac{\lambda_i}{\omega_i}\sum_{n_i=0}^{\infty}\left(\frac{\lambda_i}{\omega_i}\right)^{n_i}\prod_{k=1}^{n_i}\frac{1}{g_i(\xi,k)}. \qquad (42)$$

Hence,

$$\sum_{\mathbf{n}\in\mathbf{N}^K} \mathrm{I}_{\{n_i>0\}} g_i(\xi,n_i)\chi^{(0)}(\xi,\mathbf{n}) = \frac{\lambda_i}{\omega_i}\sum_{\mathbf{n}\in\mathbf{N}^K}\chi^{(0)}(\xi,\mathbf{n}). \tag{43}$$

It follows from (36), (37) and (41), that

$$\omega_0\frac{\partial}{\partial\xi}\left\{\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi^{(1)}(\xi,\mathbf{n})\right\}+\lambda_0\left(\frac{\mathrm{d}X_0}{\mathrm{d}\xi}+\frac{\mathrm{d}^2X_0}{\mathrm{d}\xi^2}\right)=0. \tag{44}$$

But, from (33) and (37),

$$\begin{aligned}&\sum_{i=1}^{K}\Big\{\big[\lambda_i\chi^{(1)}(\xi,\mathbf{n})-g_i(\xi,n_i+1)\omega_i\chi^{(1)}(\xi,\mathbf{n}+\mathbf{e}_i)\big]\\&\qquad-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\chi^{(1)}(\xi,\mathbf{n}-\mathbf{e}_i)-g_i(\xi,n_i)\omega_i\chi^{(1)}(\xi,\mathbf{n})\big]\Big\}\\&\quad=\omega_0\frac{\partial}{\partial\xi}\left\{\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi^{(0)}(\xi,\mathbf{n})\right\}.\end{aligned} \tag{45}$$

We let

$$s_l^{(1)}(\xi,n_l)=\sum_{j\neq l}\sum_{n_j=0}^{\infty}\chi^{(1)}(\xi,\mathbf{n}). \tag{46}$$

Thus, from (41), (42), (45), and (46),

$$\begin{aligned}&\lambda_l s_l^{(1)}(\xi,n_l)-g_l(\xi,n_l+1)\omega_l s_l^{(1)}(\xi,n_l+1)\\&\quad-\mathrm{I}_{\{n_l>0\}}\big[\lambda_l s_l^{(1)}(\xi,n_l-1)-g_l(\xi,n_l)\omega_l s_l^{(1)}(\xi,n_l)\big]\\&=\omega_0\frac{\partial}{\partial\xi}\left\{\left[\frac{\lambda_l}{\omega_l}-\mathrm{I}_{\{n_l>0\}}g_l(\xi,n_l)\right]C_l(\xi)\left(\frac{\lambda_l}{\omega_l}\right)^{n_l}\prod_{k=1}^{n_l}\frac{1}{g_l(\xi,k)}X_0(\xi)\right\}.\end{aligned} \tag{47}$$

If we sum over n_l from 0 to r, we obtain

$$\begin{aligned}&\lambda_l s_l^{(1)}(\xi,r)-g_l(\xi,r+1)\omega_l s_l^{(1)}(\xi,r+1)\\&\quad=\omega_0\frac{\partial}{\partial\xi}\left[C_l(\xi)\left(\frac{\lambda_l}{\omega_l}\right)^{r+1}\prod_{k=1}^{r}\frac{1}{g_l(\xi,k)}X_0(\xi)\right].\end{aligned} \tag{48}$$

From (46) and (48), we obtain

$$\begin{aligned}&\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]\chi^{(1)}(\xi,\mathbf{n})\\&\quad=\sum_{i=1}^{K}\sum_{n_i=0}^{\infty}\left[\frac{\lambda_i}{\omega_i}-\mathrm{I}_{\{n_i>0\}}g_i(\xi,n_i)\right]s_i^{(1)}(\xi,n_i)\end{aligned}$$

$$= \sum_{i=1}^{K} \sum_{n_i=0}^{\infty} \left[\frac{\lambda_i}{\omega_i} s_i^{(1)}(\xi, n_i) - g_i(\xi, n_i + 1) s_i^{(1)}(\xi, n_i + 1) \right]$$

$$= \sum_{i=1}^{K} \frac{\omega_0}{\omega_i} \sum_{r=0}^{\infty} \frac{\partial}{\partial \xi} \left[C_i(\xi) \left(\frac{\lambda_i}{\omega_i} \right)^{r+1} \prod_{k=1}^{r} \frac{1}{g_i(\xi, k)} X_0(\xi) \right]. \tag{49}$$

It follows from (40) and (44) that

$$\frac{dX_0}{d\xi} + \left[1 + \frac{\omega_0^2}{\lambda_0} \sum_{i=1}^{K} \frac{\lambda_i}{\omega_i^2} \right] \frac{\mathrm{d}^2 X_0}{\mathrm{d}\xi^2} = 0. \tag{50}$$

From the normalization condition, and the Euler–Maclaurin summation formula, $\int_0^\infty X_0(\xi)\,\mathrm{d}\xi = 1$. Hence, $X_0(\xi) = \tau \mathrm{e}^{-\tau\xi}$, where $\tau^{-1} = 1 + \frac{\omega_0^2}{\lambda_0} \sum_{i=1}^{K} \frac{\lambda_i}{\omega_i^2}$, independently of $g_i(\xi, n_i)$. This extends the result in Sect. 3, corresponding to $g_i(\xi, n_i) = \gamma_i$, a constant.

6 Non-work-conserving case

We next examine the heavy-traffic behavior when the work-conserving assumption does not hold, and the service rates μ_i and κ_i depend only on n_i, $i = 1, \ldots, K$, and not on the length n_0 of the primary queue. We assume that

$$\lambda_0 \big[1 + \delta\theta(\mathbf{n}) \big] = \mu_0(\mathbf{n}) + \sum_{i=1}^{K} \left[\kappa_i(n_i) - \frac{\lambda_i \kappa_i(n_i + 1)}{\mu_i(n_i + 1)} \right], \tag{51}$$

where λ_i, $i = 0, 1, \ldots, K$, and δ, $0 < \delta << 1$ are constants, $0 < \underline{\theta} \le \theta(\mathbf{n}) \le \overline{\theta} = \mathrm{O}(1)$, and $\mu_i(n_i)/\lambda_i \ge 1/\rho_i$ for $n_i \ge 1$, $i = 1, \ldots, K$.

We proceed essentially as in the work-conserving case, but now, with $\chi(\xi, \mathbf{n}; \delta)$ as in (32),

$$\begin{aligned}
&\sum_{i=1}^{K} \big\{ \big[\lambda_i + \mathrm{I}_{\{n_i > 0\}} \mu_i(n_i) \big] \chi(\xi, \mathbf{n}; \delta) \\
&\qquad - \mathrm{I}_{\{n_i > 0\}} \lambda_i \chi(\xi, \mathbf{n} - \mathbf{e}_i; \delta) - \mu_i(n_i + 1) \chi(\xi, \mathbf{n} + \mathbf{e}_i; \delta) \big\} \\
&\quad = \delta \sum_{i=1}^{K} \left[\frac{\lambda_i \kappa_i(n_i + 1)}{\mu_i(n_i + 1)} - \mathrm{I}_{\{n_i > 0\}} \kappa_i(n_i) \right] \frac{\partial \chi}{\partial \xi}(\xi, \mathbf{n}; \delta) \\
&\qquad + \delta^2 \lambda_0 \left[\theta(\mathbf{n}) \frac{\partial \chi}{\partial \xi}(\xi, \mathbf{n}; \delta) + \frac{\partial^2 \chi}{\partial \xi^2}(\xi, \mathbf{n}; \delta) \right] \\
&\qquad + \frac{\delta^2}{2} \sum_{i=1}^{K} \left[\frac{\lambda_i \kappa_i(n_i + 1)}{\mu_i(n_i + 1)} - \mathrm{I}_{\{n_i > 0\}} \kappa_i(n_i) \right] \frac{\partial^2 \chi}{\partial \xi^2}(\xi, \mathbf{n}; \delta) + \mathrm{O}\big(\delta^3\big).
\end{aligned} \tag{52}$$

Equation (34) still holds, but instead of (35) we have

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\mathrm{I}_{\{n_i>0\}}\mu_i(n_i)\chi(\xi,\mathbf{n}-\mathbf{e}_i;\delta)=\sum_{i=1}^{K}\sum_{\mathbf{n}\in\mathbf{N}^K}\mu_i(n_i+1)\chi(\xi,\mathbf{n}+\mathbf{e}_i;\delta). \quad (53)$$

Hence, if we sum over $\mathbf{n}$ in (52) and divide by δ, we obtain

$$\begin{aligned}&\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\frac{\partial\chi}{\partial\xi}(\xi,\mathbf{n};\delta)\\&\quad+\delta\lambda_0\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\theta(\mathbf{n})\frac{\partial\chi}{\partial\xi}(\xi,\mathbf{n};\delta)+\frac{\partial^2\chi}{\partial\xi^2}(\xi,\mathbf{n};\delta)\right]\\&\quad+\frac{\delta}{2}\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\frac{\partial^2\chi}{\partial\xi^2}(\xi,\mathbf{n};\delta)\\&\quad+\mathrm{O}\big(\delta^2\big)=0. \quad (54)\end{aligned}$$

From (52), if we expand in powers of δ as in (37), we obtain

$$\begin{aligned}&\sum_{i=1}^{K}\big\{\big[\lambda_i\chi^{(0)}(\xi,\mathbf{n})-\mu_i(n_i+1)\chi^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i)\big]\\&\quad-\mathrm{I}_{\{n_i>0\}}\big[\lambda_i\chi^{(0)}(\xi,\mathbf{n}-\mathbf{e}_i)-\mu_i(n_i)\chi^{(0)}(\xi,\mathbf{n})\big]\big\}=0. \quad (55)\end{aligned}$$

With an empty product being taken equal to 1, the stationary distribution with $\sum_{\mathbf{n}\in\mathbf{N}^K}\chi^{(0)}(\xi,\mathbf{n})=X_0(\xi)$ is

$$\chi^{(0)}(\xi,\mathbf{n})=\prod_{i=1}^{K}C_i\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}X_0(\xi), \quad (56)$$

where

$$C_i=\left[\sum_{n_i=0}^{\infty}\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}\right]^{-1}, \quad (57)$$

and now

$$\lambda_i\chi^{(0)}(\xi,\mathbf{n})=\mu_i(n_i+1)\chi^{(0)}(\xi,\mathbf{n}+\mathbf{e}_i). \quad (58)$$

We have the identity

$$\sum_{n_i=0}^{\infty}\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}=\sum_{n_i=0}^{\infty}\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}. \quad (59)$$

Hence,

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\chi^{(0)}(\xi,\mathbf{n})=0,\quad \text{for } i=1,\ldots,K. \tag{60}$$

It follows from (37), (54), (56), and (60), that

$$\begin{aligned}&\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\frac{\partial\chi^{(1)}}{\partial\xi}(\xi,\mathbf{n})\\&\quad+\lambda_0\left(\alpha\frac{\mathrm{d}X_0}{\mathrm{d}\xi}+\frac{\mathrm{d}^2X_0}{\mathrm{d}^2\xi}\right)=0,\end{aligned} \tag{61}$$

where

$$\alpha=\sum_{\mathbf{n}\in\mathbf{N}^K}\theta(\mathbf{n})\prod_{i=1}^{K}C_i\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}. \tag{62}$$

But, from (37) and (52),

$$\begin{aligned}&\sum_{i=1}^{K}\{\left[\lambda_i\chi^{(1)}(\xi,\mathbf{n})-\mu_i(n_i+1)\chi^{(1)}(\xi,\mathbf{n}+\mathbf{e}_i)\right]\\&\qquad-\mathrm{I}_{\{n_i>0\}}\left[\lambda_i\chi^{(1)}(\xi,\mathbf{n}-\mathbf{e}_i)-\mu_i(n_i)\chi^{(1)}(\xi,\mathbf{n})\right]\}\\&\quad=\sum_{i=1}^{K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\frac{\partial\chi^{(0)}}{\partial\xi}(\xi,\mathbf{n}).\end{aligned} \tag{63}$$

From (56), (59) and (63), with $s_l^{(1)}(\xi,n_l)$ defined as in (46), we find that

$$\begin{aligned}&\lambda_l s_l^{(1)}(\xi,n_l)-\mu_l(n_l+1)s_l^{(1)}(\xi,n_l+1)\\&\qquad-\mathrm{I}_{\{n_l>0\}}\left[\lambda_l s_l^{(1)}(\xi,n_l-1)-\mu_l(n_l)s_l^{(1)}(\xi,n_l)\right]\\&\quad=\left[\frac{\lambda_l\kappa_l(n_l+1)}{\mu_l(n_l+1)}-\mathrm{I}_{\{n_l>0\}}\kappa_l(n_l)\right]C_l\prod_{k=1}^{n_l}\frac{\lambda_l}{\mu_l(k)}\frac{\mathrm{d}X_0}{\mathrm{d}\xi}.\end{aligned} \tag{64}$$

If we sum over n_l from 0 to r, we obtain

$$\lambda_l s_l^{(1)}(\xi,r)-\mu_l(r+1)s_l^{(1)}(\xi,r+1)=C_l\kappa_l(r+1)\prod_{k=1}^{r+1}\frac{\lambda_l}{\mu_l(k)}\frac{\mathrm{d}X_0}{\mathrm{d}\xi}. \tag{65}$$

From (46) and (65), we obtain

$$\sum_{\mathbf{n}\in\mathbf{N}^K}\sum_{i=1}^{K}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]\chi^{(1)}(\xi,\mathbf{n})$$

$$=\sum_{i=1}^{K}\sum_{n_i=0}^{\infty}\left[\frac{\lambda_i\kappa_i(n_i+1)}{\mu_i(n_i+1)}-\mathrm{I}_{\{n_i>0\}}\kappa_i(n_i)\right]s_i^{(1)}(\xi,n_i)$$

$$=\sum_{i=1}^{K}\sum_{n_i=0}^{\infty}\frac{\kappa_i(n_i+1)}{\mu_i(n_i+1)}\left[\lambda_i s_i^{(1)}(\xi,n_i)-\mu_i(n_i+1)s_i^{(1)}(\xi,n_i+1)\right]$$

$$=\lambda_0\beta\frac{\mathrm{d}X_0}{\mathrm{d}\xi}, \tag{66}$$

where

$$\beta=\frac{1}{\lambda_0}\sum_{i=1}^{K}\lambda_i C_i\sum_{n_i=0}^{\infty}\left[\frac{\kappa_i(n_i+1)}{\mu_i(n_i+1)}\right]^2\prod_{k=1}^{n_i}\frac{\lambda_i}{\mu_i(k)}. \tag{67}$$

Hence, from (61) and (66), after an integration with respect to ξ,

$$\alpha X_0(\xi)+(1+\beta)\frac{\mathrm{d}X_0}{\mathrm{d}\xi}=0. \tag{68}$$

From the normalization condition, and the Euler–Maclaurin summation formula, $\int_{\xi=0}^{\infty}X_0(\xi)\,\mathrm{d}\xi=1$. Thus,

$$X_0(\xi)=\frac{\alpha}{1+\beta}\exp\left(-\frac{\alpha\xi}{1+\beta}\right). \tag{69}$$

In the work-conserving case, $\mu_i(n_i)=g_i(n_i)\omega_i$, $\kappa_i(n_i)=g_i(n_i)\omega_0$, $i=1,\ldots,K$,

$$\mu_0(\mathbf{n})=\left[1-\sum_{i=1}^{K}g_i(n_i)\right]\omega_0,$$

and $\theta(\mathbf{n})\equiv 1$. Hence, from (57), (62) and (67), $\alpha=1$, and $\beta=\frac{\omega_0^2}{\lambda_0}\sum_{i=1}^{K}\frac{\lambda_i}{\omega_i^2}$, consistent with the earlier result.

The more general case, where $\mu_i=\mu_i(\xi,n_i)$ and $\kappa_i=\kappa_i(\xi,n_i)$, $i=1,\ldots,K$, $\mu_0=\mu_0(\xi,\mathbf{n})$ and $\theta=\theta(\xi,\mathbf{n})$, leads to a first-order differential equation for $X_0(\xi)$ with coefficients that depend on ξ, which may be solved by quadrature. However, the stability condition that $X_0(\xi)\to 0$ as $\xi\to\infty$ is not readily expressed by assumptions on the service rates, so we omit the details.

7 Workload asymptotics

In order to support the heuristic arguments in Sect. 4, we now provide a rigorous proof of the counterparts of Properties (i)–(iii) in terms of workloads rather than

queue lengths. We will show that the (scaled) class-0 workload and all the class-i workloads, $i = 1, \ldots, K$, are asymptotically independent, and obtain the marginal distributions as side-results. For convenience, we focus on the work-conserving case, so that we may analyze an equivalent single-server system whose capacity is shared between $K+1$ traffic classes. We will, in fact, allow the generic service requirement B_i of class-i customers to have a general distribution $B_i(\cdot)$ with Laplace–Stieltjes Transform $\tilde{B}_i(s) = \mathbb{E}\{e^{-sB_i}\}$, mean $\beta_i = 1/\omega_i$ and second moment $\beta_i^{(2)} < \infty$, $i = 0, 1, \ldots, K$. As long as class 0 is backlogged, each backlogged class $i \notin E$ receives a portion γ_i of the capacity, while class 0 receives the remaining fraction $1 - \sum_{i \notin E} \gamma_i$, with $\mu_i = \gamma_i \omega_i$, $\kappa_i = \gamma_i \omega_0$, and $\mu_0 = (1 - \sum_{i=1}^{K} \gamma_i)\omega_0$. Note that $1 - \sum_{i=0}^{K} \lambda_i \beta_i = \lambda_0 \beta_0 \delta$.

Denote by V_i the class-i workload, i.e., the sum of the remaining service requirements of all class-i customers, $i = 0, 1, \ldots, K$, and further let $V := \sum_{i=0}^{K} V_i$. Denote by $V_i^{\gamma_i}$ the stationary workload in an isolated system with constant service rate $\gamma_i \geq \lambda_i \beta_i$ and class-i traffic only, $i = 1, \ldots, K$. According to the Pollaczek–Khinchine formula, the Laplace–Stieltjes Transform of $V_i^{\gamma_i}$ is given by $\mathbb{E}\{e^{-sV_i^{\gamma_i}}\} = \frac{(1-\rho_i)\gamma_i s}{\gamma_i s + \lambda_i (1 - \tilde{B}_i(s))}$, and

$$\mathbb{P}\{V_i^{\gamma_i} \leq t\} = (1-\rho_i) \sum_{n=0}^{\infty} \rho_i^n B_i^{r,n*}(t), \tag{70}$$

with $\rho_i = \lambda_i \beta_i / \gamma_i$ and $B_i^{r,n^*}(t)$ denoting the n-fold convolution of the class-i residual service requirement distribution $B_i^r(t) = \frac{1}{\beta_i} \int_{u=0}^{t} B_i(u)\,\mathrm{d}u$. Define $s_i^* := \arg\sup_{t-u \leq s \leq t} \{A_i(s,t) - \gamma(t-s)\}$. Standard heavy-traffic results imply that $\lambda_0 \beta_0 \delta V$ converges to an exponentially distributed random variable with mean $\sum_{i=0}^{K} \lambda_i \beta_i^{(2)}/2$ as $\delta \downarrow 0$, and hence δV converges to an exponentially distributed random variable with mean

$$\sigma^{-1} := \frac{\beta_0^{(2)}}{2\beta_0} + \frac{1}{2\lambda_0 \beta_0} \sum_{i=1}^{K} \lambda_i \beta_i^{(2)}. \tag{71}$$

Proposition 1

$$\mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1, \ldots, K\} \sim \mathbb{P}\{W > x\} \prod_{i=1}^{K} \mathbb{P}\{V_i^{\gamma_i} > y_i\} \tag{72}$$

as $\delta \downarrow 0$, *with*

$$\mathbb{P}\{W > x\} = e^{-\sigma x}.$$

Proof The proof consists of a lower bound and an upper bound that asymptotically agree as $\delta \downarrow 0$. We first establish the lower bound. From Lemma 1 below, we obtain for any $t, u \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1, \ldots, K\} \\
&\quad \geq \mathbb{P}\Big\{ V(t-u) > x/\delta + (1+K\epsilon)u, V_i(t-u) < \epsilon u, \\
&\qquad \sup_{t-u \leq s \leq t} \big\{ A_i(s,t) - \gamma_i(t-s) \big\} > y_i,\ i = 1, \ldots, K \Big\} \\
&\quad = \mathbb{P}\big\{ V(t-u) > x/\delta + (1+K\epsilon)u,\ V_i(t-u) < \epsilon u,\ i = 1, \ldots, K \big\} \\
&\qquad \times \prod_{i=1}^{K} \mathbb{P}\Big\{ \sup_{t-u \leq s \leq t} \big\{ A_i(s,t) - \gamma_i(t-s) \big\} > y_i \Big\} \\
&\quad \geq \mathbb{P}\big\{ V(t-u) > x/\delta + (1+K\epsilon)u,\ V_i(t-u) < \epsilon u,\ i = 1, \ldots, K \big\} \\
&\qquad \times \prod_{i=1}^{K} \mathbb{P}\big\{ V_i^{\gamma_i}(t) > y_i, s_i^* \geq t-u \big\} \\
&\quad \geq \left(\mathbb{P}\big\{ V > x/\delta + (1+\epsilon)u \big\} - \sum_{i=1}^{K} \mathbb{P}\{V_i \geq \epsilon u\} \right) \\
&\qquad \times \prod_{i=1}^{K} \big(\mathbb{P}\big\{ V_i^{\gamma_i} > y_i \big\} - \mathbb{P}\big\{ s_i^* < t-u \big\} \big),
\end{aligned}
$$

where we have used independence in the second step.

Letting $\delta \downarrow 0$, it follows that, for any $t, u \geq 0$,

$$
\begin{aligned}
&\liminf_{\delta \downarrow 0} \mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1, \ldots, K\} \\
&\quad \geq \left(\liminf_{\delta \downarrow 0} \mathbb{P}\{V > x/\delta\} - \sum_{i=1}^{K} \mathbb{P}\{V_i \geq \epsilon u\} \right) \prod_{i=1}^{K} \big(\mathbb{P}\big\{ V_i^{\gamma_i} > y_i \big\} - \mathbb{P}\big\{ s_i^* < t-u \big\} \big).
\end{aligned}
$$

Since we may choose the value of u arbitrarily large, we deduce

$$
\begin{aligned}
&\liminf_{\delta \downarrow 0} \mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1, \ldots, K\} \\
&\quad \geq \liminf_{\delta \downarrow 0} \mathbb{P}\{\delta V > x\} \prod_{i=1}^{K} \mathbb{P}\big\{ V_i^{\gamma_i} > y_i \big\} = \mathrm{e}^{-\sigma x} \prod_{i=1}^{K} \mathbb{P}\big\{ V_i^{\gamma_i} > y_i \big\}.
\end{aligned}
$$

We now prove the upper bound. From Lemma 2 below, we obtain, for any $t, u \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1, \ldots, K\} \\
&\quad \leq \big(\mathbb{P}\big\{ A_0(t-u,t) > (\rho_0 + \epsilon)u \big\} + \mathbb{P}\big\{ V_0(t-u) > x/\delta - (\rho_0 + \epsilon)u \big\} \big) \\
&\qquad \times \prod_{i=1}^{K} \big(\mathbb{P}\big\{ s_i^* < t-u \big\} + \mathbb{P}\big\{ V_i^{\gamma_i}(t) > y_i \big\} \big)
\end{aligned}
$$

$$\leq \big(\mathbb{P}\big\{A_0(t-u,t) > (\rho_0+\epsilon)u\big\} + \mathbb{P}\big\{V_0 > x/\delta - (\rho_0+\epsilon)u\big\}\big)$$
$$\times \prod_{i=1}^{K}\big(\mathbb{P}\big\{s_i^* < t-u\big\} + \mathbb{P}\big\{V_i^{\gamma_i} > y_i\big\}\big),$$

where we have used independence.

Letting $\delta \downarrow 0$, it follows that, for any $t, u \geq 0$,

$$\limsup_{\delta\downarrow 0} \mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1,\ldots,K\}$$
$$\leq \Big(\mathbb{P}\big\{A_0(t-u,t) > (\rho_0+\epsilon)u\big\} + \limsup_{\delta\downarrow 0}\mathbb{P}\{\delta V > x\}\Big)$$
$$\times \prod_{i=1}^{K}\big(\mathbb{P}\big\{s_i^* < t-u\big\} + \mathbb{P}\big\{V_i^{\gamma_i} > y_i\big\}\big).$$

Since we may choose the value of u arbitrarily large, we deduce

$$\limsup_{\delta\downarrow 0} \mathbb{P}\{\delta V_0 > x,\ V_i > y_i,\ i = 1,\ldots,K\}$$
$$\leq \limsup_{\delta\downarrow 0}\mathbb{P}\{\delta V > x\}\prod_{i=1}^{K}\mathbb{P}\big\{V_i^{\gamma_i} > y_i\big\} = \mathrm{e}^{-\sigma x}\prod_{i=1}^{K}\mathbb{P}\big\{V_i^{\gamma_i} > y_i\big\}. \qquad \square$$

We now prove the two sample-path relationships that were used in the proof of Proposition 1. Denote by $V_i(t)$ the amount of class-i work in the system at time t, $i = 0, 1, \ldots, K$, and let $V(t) := \sum_{i=0}^{K} V_i(t)$ be the total amount of work in the system at time t. For any $s \leq t$, denote by $A_i(s,t)$ and $B_i(s,t)$ the amount of traffic generated and the amount of service received by class i during the time interval $[s,t]$, respectively. By definition, $V_i(r,s) = V_i(r) + A_i(r,s) - B_i(r,s)$ for all $r \leq s$.

Lemma 1 *For any $t, u \geq 0$, if $V(t-u) > x + (1+K\epsilon)u$, $V_i(t-u) < \epsilon u$, $\sup_{t-u\leq s\leq t}\{A_i(s,t) - \gamma_i(t-s)\} > y_i$, $i = 1,\ldots,K$, then $V_0(t) > x$, $V_i(t) > y_i$, $i = 1,\ldots,K$.*

Proof We have that $\sum_{i=0}^{K} B_i(r,s) \leq s - r$ for all $r \leq s$. Thus, $V_0(s) = V_0(t-u) + A_0(t-u,s) - B_0(t-u,s) \geq V(t-u) - \sum_{i=1}^{K} V_i(t-u) - \sum_{i=0}^{K} B_i(t-u,s) > x + u - (s-(t-u)) = x + t - s \geq 0$ for all $s \in [t-u,t]$, and in particular $V_0(t) \geq x$. The above also implies $B_i(s,t) \geq \gamma_i(t-s)$, $i = 1,\ldots,K$, for all $s \in [t-u,t]$. Noting that $A_i(s^*,t) - \gamma_i(t-s_i^*) > y$, it then follows that $V_i(t) = V_i(s_i^*) + A_i(s_i^*,t) - B_i(s_i^*,t) \geq A_i(s_i^*,t) - \gamma_i(t-s_i^*) > y_i$. $\square$

Lemma 2 *For any $t, u \geq 0$, if $V_0(t) > x$, $V_i(t) > y_i$, $i = 1,\ldots,K$, then either $A_0(t-u,t) > (\rho_0+\epsilon)u$ or $V_0(t-u) > x - (\rho_0+\epsilon)u$ and either $s_i^* < t-u$ or $\sup_{t-u\leq s\leq t}\{A_i(s,t) - \gamma_i(t-s)\} > y_i$, $i = 1,\ldots,K$.*

Proof Suppose not. If $A_0(t-u,t) \leq (\rho_0+\epsilon)u$ and $V_0(t-u) \leq x-(\rho_0+\epsilon)u$, then $V_0(t) = V_0(t-u) + A_0(t-u,t) - B_0(t-u,t) \leq x$. Otherwise, if $s_i^* \geq t-u$ and $\sup_{t-u\leq s\leq t}\{A_i(s,t)-\gamma_i(t-s)\} \leq y_i$, then $V_i(t) \leq V_i^{\gamma_i}(t) = \sup_{s\leq t}\{A_i(s,t)-\gamma_i(t-s)\} = \sup_{t-u\leq s\leq t}\{A_i(s,t)-\gamma_i(t-s)\} \leq y_i$. Hence, in both cases a contradiction arises. □

8 Conclusion

We have analyzed a system of parallel queues which are coupled through their service rates, causing a complex dynamic interaction. Specifically, the system consists of one primary queue and several secondary queues whose service rates depend on whether the primary queue is empty or not. Conversely, the service rate of the primary queue depends on which of the secondary queues are empty. We considered a heavy-traffic scenario where each of the secondary queues is assumed to be underloaded when the primary queue is busy.

In Sects. 2 and 3 we used a perturbation approach to derive the lowest-order asymptotic approximation to the joint stationary distribution of the queue lengths, in terms of a small positive parameter measuring the closeness of the system to instability. These results were extended to more general scenarios in Sects. 5 and 6. In Sects. 4 and 7 we presented heuristic derivations of these results based on probabilistic arguments, which were further supported by a sample path approach to establish corresponding joint asymptotics of the workloads.

It is interesting to compare these two approaches. The analytical perturbation approach relies on a set of balance equations for the stationary probabilities of a Markov process, which requires both Poisson arrivals and exponentially distributed service requirements, and is better-suited to queue lengths. The queue lengths, however, can be easily translated into workloads, and this yields joint workload asymptotics that agree with those obtained using the sample-path approach. Also, the approach extends to scenarios where the service rates depend on the exact queue lengths, and not just on whether a queue is empty or not.

In contrast, the sample-path approach involves probabilistic lower and upper bounds, which allow for generally distributed service requirements (as well as non-Poisson arrivals with some additional effort), and is more convenient for workloads. In their turn, however, the workloads can be mapped to queue lengths, and this allows extension of the joint queue length asymptotics to generally distributed service requirements. As a limitation though, the approach is more difficult to extend to scenarios where the service rates do not just depend on whether a queue is empty or not, but also on the exact queue lengths.

Acknowledgements The authors are grateful to the anonymous referees for their helpful comments and suggestions.

References

1. Bertsimas, D., Paschalidis, I.Ch., Tsitsiklis, J.N.: Large deviations analysis of the generalized processor sharing policy. Queueing Syst. **32**, 319–349 (1999)
2. Borst, S.C., Boxma, O.J., Jelenković, P.R.: Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. Queueing Syst. **43**, 273–306 (2003)
3. Borst, S.C., Boxma, O.J., Van Uitert, M.J.G.: The asymptotic workload behavior of two coupled queues. Queueing Syst. **43**, 81–102 (2003)
4. Borst, S.C., Mandjes, M.R.H., Van Uitert, M.J.G.: Generalized processor sharing queues with heterogeneous traffic classes. Adv. Appl. Probab. **35**, 806–845 (2003)
5. Cohen, J.W., Boxma, O.J.: Boundary Value Problems in Queueing System Analysis. North-Holland, Amsterdam (1983)
6. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Probab. **5**, 49–77 (1995)
7. Elwalid, A., Mitra, D.: Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes. In: Proc. IEEE Infocom '99, New York, USA, pp. 1220–1230 (1999)
8. Fayolle, G., Iasnogorodski, R.: Two coupled processors: the reduction to a Riemann–Hilbert problem. Z. Wahrscheinlichkeitstheor. Verw. Geb. **47**, 325–351 (1979)
9. Guillemin, F., Mazumdar, R.R., Dupuis, A., Boyer, J.: Analysis of the fluid weighted fair queueing system. J. Appl. Probab. **40**, 180–199 (2003)
10. Guillemin, F., Pinchon, D.: Analysis of generalized processor-sharing systems with two classes of customers and exponential services. J. Appl. Probab. **41**, 832–858 (2004)
11. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley, Chichester (1979)
12. Knessl, C., Morrison, J.A.: Heavy-traffic analysis of two coupled processors. Queueing Syst. **43**, 173–220 (2003)
13. Konheim, A.G., Meilijson, I., Melkman, A.: Processor sharing of two parallel lines. J. Appl. Probab. **18**, 952–956 (1981)
14. Kumaran, K., Margrave, G., Mitra, D., Stanley, K.: Novel techniques for the design and control of generalized processor sharing schedulers for multiple QoS classes. In: Proc. IEEE Infocom 2000, Tel-Aviv, Israel (2000)
15. Massoulié, L.: Large deviations estimates for polling and weighted fair queueing service systems. Adv. Perf. Anal. **2**, 103–128 (1999)
16. Morrison, J.A.: Diffusion approximation for head-of-the-line processor sharing for two parallel queues. SIAM J. Appl. Math. **53**, 471–490 (1993)
17. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM Trans. Netw. **1**, 344–357 (1993)
18. Parekh, A.K., Gallager, R.G.: A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. IEEE/ACM Trans. Netw. **2**, 137–150 (1994)
19. Ramanan, K., Reiman, M.I.: Fluid and heavy-traffic diffusion limits for a generalized processor sharing model. Ann. Appl. Probab. **13**, 100–139 (2003)
20. Ramanan, K., Reiman, M.I.: The heavy-traffic limit of an unbalanced generalized processor sharing model. Ann. Appl. Probab. **18**, 22–58 (2008)
21. Yaron, O., Sidi, M.: Generalized Processor Sharing networks with exponentially bounded burstiness arrivals. In: Proc. IEEE Infocom '94, pp. 628–634 (1994)
22. Zhang, Z.-L.: Large deviations and the generalized processor sharing scheduling for a multiple-queue system. Queueing Syst. **28**, 349–376 (1998)